Neurocomputing 149 (2015) 641–651

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Support vector machine with parameter optimization by a novel hybrid method and its application to fault diagnosis

Xiaoyuan Zhang [a,*], Daoyin Qiu [b,c], Fuan Chen [a]

[a] College of Electrical Engineering, Henan University of Technology, Zhengzhou 450001, PR China
[b] Huazhong University of Science and Technology, Wuhan 430074, PR China
[c] North China University of Water Conservancy and Hydropower, PR China

ARTICLE INFO

Article history:
Received 26 March 2014
Received in revised form
3 July 2014
Accepted 4 August 2014
Communicated by J. Zhang
Available online 19 August 2014

Keywords:
Support vector machine
Parameter optimization
Differential evolution
Particle swarm optimization
Bare bones differential evolution
Fault diagnosis

ABSTRACT

The performance of support vector machine (SVM) heavily depends on its parameters. The parameter optimization for SVM is still an ongoing research issue. The current parameter optimization methods either are easy to fall into local optimal solution, or are time consuming. Moreover, some optimization methods depend also on the choice of parameters for them, provoking thus a vicious circle. In view of this, a new hybrid method is proposed to optimize the parameters of SVM in this paper. It uses the inter-cluster distance in the feature space (ICDF) to determine a small and effective search interval from a larger kernel parameter search space, while a hybrid of the barebones particle swarm optimization and differential evolution (BBDE) is used to search the optimal parameter combination in the new search space. The ICDF shows the degree the classes are separated. The BBDE is a new, almost parameter-free optimization algorithm. Some benchmark datasets are used to evaluate the proposed algorithm. Furthermore, the proposed method is used to diagnose the faults of rolling element bearings. Experiments and engineering application show that the proposed method outperforms other methods both mentioned in this paper and published in other literature.

1. Introduction

Nowadays, with the economics pressure to reduce the costs, the downtime of industrial plants and to shorten the time necessary to manufacture a product, the fault diagnosis (FD) plays an important role. FD makes it possible to determine the place, reason and time of the fault precisely. An effective and efficient FD method is an important issue most techniques have sought to. As a state-of-the-art pattern recognition method, support vector machine (SVM) has been vastly used in FD [1–5]. The theory of SVM is based on the structural risk minimization (SRM) principle [6]. It has a good performance in solving nonlinear and high dimensional pattern recognition problems with small-sample size. However, poor choice of parameter setting can dramatically decrease the generalization performance of SVM. The parameters of SVM includes the value of the regularization constant C and the Kernel type, with its respective parameters. Up to now, there is no systematic methodology or priori knowledge for determining the parameters of SVM. Thus, it will be desirable to have an effective and automatic model selection scheme to make SVM practical for real life applications, in particular, for people who are not familiar with parameters tuning in SVM [7].

Many studies have been carried out on SVM parameters optimization. A straightforward way is the simple exhaustive grid search (GS) in the parameter space [8]. GS trains SVMs with all desired combinations of parameters and screens them according to the training accuracy. It makes the training process time-consuming, and when the number of parameters exceeds two it will become intractable. Furthermore, to reach high performances, GS needs a good discretization of the search space in fixed values, which is a tricky task since a suitable sampling step varies from kernel to kernel and the grid interval may not be easy to locate without prior knowledge of the problem [9]. Therefore, many improved parameters selection methods have been proposed. Chapelle et al. [10,11] used gradient-based numerical optimization methods to find the optimal parameters for SVM. By minimizing some generalization bounds such as the leave-one-out (LOO) error bounds, the numerical methods are generally more efficient than GS. There are also other numerical optimization approaches in the literature [12–16]. However, owing to the non-convexity of the generalization bounds, these methods may get stuck into local optimum and cause instabilities. Moreover, these methods are sensitive to the initial point and require that the kernel functions and the bounds of generalization error have to be differentiable with respect to kernel and penalty constant parameters. To

* Corresponding author. Tel.: +86 18625531521.
E-mail address: xyzhang27@gmail.com (X. Zhang).

http://dx.doi.org/10.1016/j.neucom.2014.08.010
0925-2312/

overcome the above limitations, some intelligent evolutionary algorithms (EA), such as genetic algorithm (GA) [17–21], covariance matrix adaptation evolution strategy (CMA-ES) [22], particle swarm optimization (PSO) [23–26], artificial immunization algorithm (AIA) [4] and ant colony optimization algorithm (ACO) [27,28] have been used to optimize the SVM parameters for their better global search abilities. EAs do not need additional information of the problem, and are generally good at finding "acceptable good" or near-optimal solutions to problems. But they are not well suited to perform finely tuned search and thus not guaranteed to find the global optimum solution to a problem. Furthermore, EAs select the best parameter combination from the population evolved generation by generation, which requires re-training SVMs many times. Thus, they are still time-consuming, especially when the search space of the parameters is large. Moreover, the performance of most EAs depend also on the choice of parameters for them, such as, for instance, the inertia weight in PSO, provoking thus a vicious circle.

From the aforementioned literature review, it can be seen that the presence of local minima in the search space, the computational time required for model selection task and the vicious circle in SVM parameter optimization process have been considered the main challenges in the field. In view of this, a novel SVM parameter optimization algorithm is presented in this paper. In the proposed method, the range of the kernel parameter is first shortened by the use of an inter-cluster distance in the feature space (ICDF) measure heuristic, which can reduce the calculation amount of the parameter optimization process and improve the chances to find the optimal parameters. Afterwards, an almost parameter-free optimization method, bare bones differential evolution (BBDE), is used to optimize the parameter values. BBDE is very powerful yet almost no parameter to be tuned. It can avoid of the vicious circle of traditional EAs.

In the following part of the introduction, we briefly review the ICDF index and the BBDE algorithm. ICDF indicates the degree the classes are separated in the feature space. When kernel function is fixed, selecting kernel parameters is equivalent to selecting the feature space. A feature space in which the classes are more separated corresponds to a better kernel parameter value for SVM. In literature [29,30], Wu and Wang used ICDF as the separation index to choose the kernel parameters for SVM. In their study, for binary classification problem, the optimal kernel parameter is determined according to the largest value of ICDF. Then, the selected kernel parameter with different candidate penalty parameters are used to train SVM models. The parameters combination which results in a highest cross validation accuracy is selected as the best one. Since calculating ICDFs does not require the information of the trained SVMs, the time needed for the training process for different kernel parameters is saved. This method can get the parameter combinations which result in SVM models perform as good as the models chosen by the traditional GS method in testing accuracy, while the training time is shortened largely. However, in their method, the candidate penalty parameters and kernel parameters for calculating ICDFs are given as discretization. It needs to locate the interval of feasible solution and a suitable sampling step, which is a tricky task. Furthermore, in practice, most of the classification problems are multi-class, how to extend the binary SVM with parameter optimization by ICDF to multi-class problems is a challenging task.

Bare bones differential evolution (BBDE), proposed by Omran et al. [31], is a hybrid of the barebones particle swarm optimizer (PSO) [32] and differential evolution (DE) [33]. It is an almost parameter-free, self-adaptive, optimization algorithm. BBDE capitalizes on the strengths of both the barebones PSO and self-adaptive DE. The BBDE does not make use of the standard PSO parameters (i.e. inertia weight, acceleration coefficients, and velocity clamping), and also removes the DE scale parameter. The only parameter is the probability of recombination. This is a highly favorable characteristic for parameter optimization of SVM, as it can overcome the aforementioned vicious circle.

The proposed method makes full use of ICDF measure and BBDE algorithm, in which ICDF measure is responsible to determine a small and effective search interval both for binary classification problem and multi-class classification problem, from a larger kernel parameter search space, then BBDE is used to search the optimal parameter combination of SVM in the continuous intervals of kernel parameter and penalty parameter. That is, the exhaustive search of parameters can be performed in the priority area indicated by ICDF measure. The search range of kernel parameter in the proposed method is much smaller yet more effective than that of traditional methods. Moreover, BBDE has shown performance superior to that of PSO and other EAs in the widely used benchmark problems [31] with almost no parameters being set. The proposed method is tested on several benchmark datasets as well as fault diagnosis for rolling element bearings. Experiments and engineering application show that the proposed method outperforms other methods both mentioned in this paper and published in other literature.

The remaining of this paper is organized as follows. The brief introduction of SVM and its parameter selection are presented in Section 2. The proposed method is described in detail in Section 3. In Section 4 the proposed method and some other optimization methods are applied in some benchmark datasets and the experimental results are compared. In Section 5 the proposed method is applied in FD for rolling element bearings. Finally, the conclusion is drawn in Section 6.

2. Support vector machine and its parameter selection

Support vector machine (SVM) is a statistical classification method based on the structural risk minimization (SRM) approach proposed by Vapnik et al. [6]. Given a dataset with n examples (x_i, y_i), where each x_i is an input data and $y_i \in \{+1,1\}$ corresponds to its bipolar label, $i=1, 2, \ldots, n$. By using a nonlinear mapping $\phi(x)$, the input data is mapped into a high dimensional feature space F, in which the data are sparse and possibly more separable. Then, the maximum margin separating hyperplane $w\phi(x)+b=0$is built in F, where w is a weight vector orthogonal to the hyperplane, b is an offset term. The margin is $1/\|w\|$. Maximizing the margin $1/\|w\|$ is equivalent to minimizing $\|w\|^2$, whose solution is found after resolving the following quadratic optimization problem:

$$\min_{\omega,b} \frac{1}{2}\left|\left|w\right|\right|^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{Restrictions}: \; y_i(w\phi(x_i)+b) \geq 1-\xi_i \text{ and } \xi_i \geq 0 \text{ for } i=1,\ldots,n \quad (1)$$

where C is the penalty parameter which imposes a trade-off between training error and generalization, ξ_i are slack variables. The restrictions are imposed to ensure that no training data should be within the margins. Nevertheless, they are relaxed by the slack variables to avoid an overfitting to noisy data. The number of training errors and examples within the margins allowed by the introduction of these variables is controlled by the minimization of the term $\sum_{i=1}^{n}\xi_i$. By using the duality theory of optimization theory, the final decision function can be given by

$$\begin{aligned} f(x) &= \text{sgn}(\sum_{x_i\in SVs} y_i a_i(\phi(x_i)\phi(x))+b) \\ &= \text{sgn}(\sum_{x_i\in SVs} y_i \alpha_i K(x_i,x)+b) \end{aligned} \quad (2)$$

where the constants a_i are called Lagrange multipliers and are determined in the optimization process. *SVs* corresponds to the set of support vectors, training examples for which the associated